UNITED STATES PATENT APPLICATION

FOR

METHOD, SYSTEM AND MACHINE READABLE MEDIUM FOR

PUBLISHING DOCUMENTS USING AN ONTOLOGICAL MODELING

SYSTEM

INVENTORS

Dan Adamson
Leo Shih
Alain T. Rappaport

Prepared by:

Blakely, Sokoloff, Taylor & Zafman
12400 Wilshire Boulevard
Seventh Floor
Los Angeles, California 90025
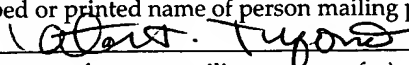(408) 720-8300

Attorney's Docket No. 004239.P006

RELATED APPLICATIONS

[0001]   This application claims the benefit of U.S. Provisional Application

No. 60/411,902, filed September 19, 2002, which is incorporated herein in its

entirety by reference.

FIELD OF THE INVENTION

[0002]   This invention relates generally to the field of information

processing and to information modeling systems. More particularly, this

invention relates to a method for publishing documents based on underlying

ontological models representative of the concepts contained within the

documents.

BACKGROUND OF THE INVENTION

[0003]   Many approaches exist to cataloging or indexing documents and

making them available as part of a collection to users. Simple search systems, for

example, allow for the selection of documents based on text (keyword) searching.

[0004]   Other approaches may involve the use of taxonomy-based

methodologies to classify documents into certain categories. In these

approaches, taxonomies of the domain are defined and documents are tagged or

sorted into categories in accordance with the relevance of these documents to

elements of the defined taxonomies. While this is useful for organizing or

classifying a collection of documents, it does little to increase their searchability if

they are not sorted using categories appropriate to the search context being used.

Furthermore, this does little to improve the ability of a user or system being able

to perform analysis on the documents.

5

SUMMARY OF THE INVENTION

[0005]   A method and system to process one or more documents in a domain. The method includes modeling the domain with a plurality of domain models using an ontological system; representing each document as a collection of one or more domain models; and populating the domain models that are used to represent the document with values corresponding to properties of the document being represented.

[0006]   The invention also extends to a machine-readable medium embodying a sequence of instructions that, when executed by a machine, cause the machine to perform anyone of the methods described herein.

[0007]   Other features of the present invention will be apparent from the accompanying drawings and from the detailed description that follows.

BREIF DESCRIPTION OF THE DRAWINGS

[0008]   The present invention is illustrated by way of example and not limitation in the figures of the accompanying drawings, in which like references indicate similar elements and in which:

[0009]   Fig. 1 is a block diagram, according to an exemplary embodiment of the present invention, illustrating an overview of a general process for utilizing an ontological modeling system for publishing documents;

[0010]   Fig. 2 is a flow chart, according to an exemplary embodiment of the present invention, illustrating a process to automatically classify documents and transform them into models;

[0011]   Fig. 3 is a block diagram, according to an exemplary embodiment of the present invention, illustrating an example implementation of the invention; and

[0012]   Fig. 4 is a diagrammatic representation of a machine in the exemplary form of a computer system within which a set of instructions, for causing the machine to perform any one or more of the methodologies discussed herein, may be executed.

DETAILED DESCRIPTION

[0013]   As discussed herein, the present invention provides a method, system, and machine-readable medium for establishing a consistent infrastructure for data modeling and publishing based on underlying ontological models. In the following description, for purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the present invention. It will be evident, however, to one skilled in the art that the present invention may be practiced without these specific details.

[0014]   The following discussion assumes an ontological or object-orientated modeling system is being used such as that described in the U.S. Provisional Application No. 60/361,746 filed on March 4, 2002, which is incorporated herein in its entirety by reference. However, alternative-modeling systems may also be used.

[0015]   In the process described herein, ontologies are used as a basis for building repositories of structured information on complex domains including, for example, molecular structures, genes or diseases. Life sciences, medical or other scientific documents are used to derive information which is entered or catalogued in the ontology. As an example, ontologies have been built to catalog detailed information on biological concepts such as genes or proteins, molecular pathways or medical conditions.

[0016]   This invention makes a novel use of ontologies, to create, for any

document (where a document may be, for example, a text, a database report,

etc.), a corresponding model of that document built by extracting relevant

elements of that text corresponding to models in an ontology. Hence, a unique

5   model is created for each individual document, which may then be used by users

or computer applications. This ontological model constitutes a "computation-

ready" version of the original document, where information is highly structured

and numerical information is maintained as part of the model. These structures

and data in the model can easily be used or exploited by any other computer

10   application. Further, when generating a model of the original document, the

ontologies may add relevant information that is not in the document itself but

augments its information content or allows a user or computer application to

augment this information.

[0017]   Traditional use of ontologies consists of taking a knowledge-base of

15   information and creating a single collective view or representation of this

knowledge as an ontology with a single consistent view. The concepts,

processes, and methodology described herein, in contrast, allow for the creation

of a collection of models based on different knowledge sources involving a loose

ontology that doesn't require consistency across all data sources. As an example,

20   one knowledge source describes a gene as a tumor suppressor gene while

another data source describes the same gene as non-functional. In one

embodiment of the invention described herein, both sources of data can be

successfully modeled independently, using the same ontology without conflict.

[0018]		According to one aspect of the present invention, an approach

discussed here involves ontological modeling in which documents are modeled

by describing them as a set of one or more instances of models. The set of

models describing the documents can define specific relationships or contexts

between the models, as described in the documents, but in a manner that is

explicit and allows for machine processing. Scientific documents are particularly

rich in context-sensitive information that is difficult to analyze or discover

without intelligent processes or without an indexing process that properly

prepares the data for further analysis. In one embodiment of the invention, a

process described herein makes search and data analysis of scientific documents

more amenable to computer and human analysts alike.

[0019]		An overview of a document publishing system is shown in **Fig. 1,**

according to an exemplary embodiment of the present invention. In one

embodiment, the documents are analyzed to create a model representation of the

documents. This process may be completed by human input or may be

completed automatically by a computer system. In one embodiment it may

involve a combination of both human and computer input, perhaps with an

initial automated analysis followed by a manual revision process.

[0020]   The model representation formed from the document is stored by

the modeling system, possibly as part of a collection with other model

representations representing other documents. In one embodiment, this

collection of modeled documents is then made searchable by various model

properties using the underlying search mechanism of the modeling system.

[0021]   In one embodiment, the system performing search and retrieval of

documents allows the requesting system the ability to retrieve models and/or the

original documents for further analysis. The model form of the document should

be understood by the querying system, such that the returned models can then

be further analyzed for specific properties. In addition, the collection could be

statistically analyzed for collective properties (e.g. averages,

minimum/maximum values, clustering analysis, etc.).

[0022]   As an example, a collection of documents describing clinical trials

could be indexed using this approach, with models being created that describe

the documents' contents, for example, with 'clinical trial' models. The clinical

trial models might include a 'number of patients' property that indicates how

many patients were involved in the clinical trial. As an example, a querying

system would then be able to search for all documents that described clinical

trials with over 50 patients. Also, after a set of models was returned from a query, the querying system could then use the 'number of patients' property to estimate the total cost of each trial, or it could use the property for each model to determine an average number of patients involved in the clinical trials.

[0023]   A sample process for automatically generating models based on keywords is shown in **Fig. 2,** according to an exemplary embodiment of the present invention. In one embodiment using this process, a collection of keywords that relate to specific objects are compared to one or more documents being analyzed. As an example, the keyword 'Aspirin' might correspond to a 'drug' object. In this case, if the word 'Aspirin' appears in a document, a 'drug' object might be created and would be populated, for example, with 'aspirin' as the common name of the drug. In more complicated embodiments, the model could be further populated, for example, with the drug's generic name ("acetylsalicylic acid") in the generic name property of the drug model.

[0024]   Initially the document is searched for keywords that appear in a pre-set vocabulary list (at block 2.1). This list could correspond to a list created manually or from a pre-existing set of populated models, or even from a set of model definitions that have, for given properties, pre-defined values. Alternatively, or in addition, the list may be created utilizing a pre-existing vocabulary/terminology system. For example, in one embodiment for health

care related terms, the MESH terminology can be easily adapted to correspond to appropriate objects based on the placement of the term within the MESH hierarchy. This approach allows for a large vocabulary with little effort.

[0025]   If a match is found with a keyword (at block 2.2), and the keyword corresponds to only one model type (block 2.3) then a model for that type would be created for that concept (at block 2.4) and stored as part of that documents model set.

[0026]   In more complicated cases, more than one keyword may be necessary to trigger the creation of an object, either due to ambiguous terminology or due to models corresponding to broader concepts that involve multiple keywords. In one case, as an example, the keyword 'HER-2' may be a reference to a 'gene' object or it may be a reference to a 'protein' object. In this case, the context of the keyword in the document differentiates between which object is created. Such a process is shown at block 2.5, where additional keywords are searched for to differentiate the meaning of the first 'trigger' keyword. If any keywords are found which enable the clarification of the context, the appropriate model object is created (block 2.6). If not, in one embodiment, the keyword could be marked for review by a human auditor of the indexing process, who could then decide which model object to create.

[0027]   Even in this process, a review by a human may still be necessary. Furthermore, while this process may create a collection of objects, this collection may not correctly model the document's description of the relationships between the objects. For example, the process may create a 'drug' object, a 'treatment' object and a 'dosage' object for a document describing a specific treatment. However, after this process it would be left to another process (e.g., human or automated, block 2.7) to insert the drug object and dosage object inside the treatment object based on the context of the document. This secondary process could be aided by tools which allowed 'dragging and dropping' of one model into another model by a human reviewer.

[0028]   In one embodiment of the invention, the system as described is used for managing resources that are typically managed by an Enterprise Resource Planning (ERP) system. In this case, the invention would be used to manipulate (store/search/modify/compute against) models that represent the resources being managed. In this fashion, deep domain-specific models could be managed.

Example Implementation

[0029]   The following example is provided for illustration and explanation

purposes as a description of one possible instance of the process described. It

should not be construed to limit the scope of the invention in anyway.

5   [0030]   In this example implementation, a group or individual wish to take

a set of publications (for example, conference or journal abstracts) that describe

clinical studies in breast cancer, and make them searchable by a variety of criteria

and perform certain analysis on them.

[0031]   Fig. 3 is a flow chart, according to an exemplary embodiment of the

10   present invention, that illustrates a process utilized by a model designer to create

models for the breast cancer domain (block 3.1). During this process, the

designer creates models for the condition, treatment modes, drugs, diagnostics

and other notions involved in the field of breast cancer. Models of the breast

cancer condition may include, for example, a property that represents the stage

15   of disease progression. As another example, patient models may have properties

that represent age, menopausal status, and may itself contain other created

models, such as a condition model. Other models might be created such as a

model of a clinical trial, with properties that allow the precise description of

different arms of the trial and the results of the trial.

[0032]   An application is then created that allows a reviewer (either human or machine) to describe the publications about breast cancer using the models. For each publication (block 3.2), the reviewer may create collections of model instances that describe the contents of that abstract (block 3.3). In this example, the reviewer may be a human who uses a web form to input information about the abstract, which is then translated by the application to create model instances.

[0033]   Each abstract or full publication is then represented by a collection of models. The model instances represent a "computation-ready" form of the original information that can be understood by other computer applications. If the above process is done for many publications, then many such collections exist, and this set of collections can then be searched for papers that match very specific criteria against information that has been captured by the models (block 3.4).

[0034]   Alternatively or in addition to searching, the set of collections of models can be analyzed for content, either individually (representing one paper) or as a set (representing the set of papers). Continuing with the example, a physician might be able to perform a search for all papers that discuss a specific disease state, and may then be able to determine using algorithms acting directly on the collection of models very specific information about the returned

collection, such as an average success rate or a ratio comparing study success rates to the study's endpoint. The specific computations allowed by a rich model set are practically limitless, and these analyses become trivial if the models include properties to capture the data pertinent for solving the problem at hand.

[0035] In one embodiment, instances from a base set of models or schemas are used to describe the document. The process, however, may itself result in new models being derived from the resulting collection of models, possibly becoming reference models as well.

[0036] Fig. 4 illustrates a diagrammatic representation of machine in the exemplary form of a computer system 300 within which a set of instructions, for causing the machine to perform any one or more of the methodologies discussed herein, may be executed. In alternative embodiments, the machine operate as a standalone device or may be connected (e.g., networked) to other machines. In a networked deployment, the machine may operate in the capacity of a server or a client machine in server-client network environment, or as a peer machine in a peer-to-peer (or distributed) network environment. The machine may be a personal computer (PC), a tablet PC, a set-top box (STB), a Personal Digital Assistant (PDA), a cellular telephone, a web appliance, a network router, switch or bridge, or any machine capable of executing a set of instructions (sequential or otherwise) that specify actions to be taken by that machine. Further, while only a

single machine is illustrated, the term "machine" shall also be taken to include any collection of machines that individually or jointly execute a set (or multiple sets) of instructions to perform any one or more of the methodologies discussed herein.

5   [0037]   The exemplary computer system 300 includes a processor 302 (e.g., a central processing unit (CPU) a graphics processing unit (GPU) or both), a main memory 304 and a static memory 306, which communicate with each other via a bus 308. The computer system 300 may further include a video display unit 310 (e.g., a liquid crystal display (LCD) or a cathode ray tube (CRT)). The

10   computer system 300 also includes an alpha-numeric input device 312 (e.g., a keyboard), a cursor control device 314 (e.g., a mouse), a disk drive unit 316, a signal generation device 318 (e.g., a speaker) and a network interface device 320. The disk drive unit 316 includes a machine-readable medium 322 on which is stored one or more sets of instructions (e.g., software 324) embodying any one or

15   more of the methodologies or functions described herein. The software 324 may also reside, completely or at least partially, within the main memory 304 and/or within the processor 302 during execution thereof by the computer system 300, the main memory 304 and the processor 302 also constituting machine-readable media.

[0038]   The software 324 may further be transmitted or received over a

network 326 via the network interface device 320.

[0039]   While the machine-readable medium 322 is shown in an exemplary

embodiment to be a single medium, the term "machine-readable medium"

should be taken to include a single medium or multiple media (e.g., a centralized

or distributed database, and/or associated caches and servers) that store the one

or more sets of instructions. The term "machine-readable medium" shall also be

taken to include any medium that is capable of storing, encoding or carrying a

set of instructions for execution by the machine and that cause the machine to

perform any one or more of the methodologies of the present invention. The

term "machine-readable medium" shall accordingly be taken to included, but not

be limited to, solid-state memories, optical and magnetic media, and carrier wave

signals.

Alternative and equivalent processes

[0040]   Thus a method, system, and machine-readable medium for

establishing a consistent infrastructure for data modeling and publishing based

on underlying ontological models have been described. Although embodiments

of the present invention have been shown and described, along with certain

variants of the invention, it should be understood and recognized by those

skilled in the art that many other varied embodiments that incorporate the

teachings of the present invention may be implemented or constructed.

Accordingly, the scope of the present invention is not to be limited to the specific

embodiments, forms or examples described herein.

5